Speech as a Feedback Modality for Smart Objects

Clemens Lombriser*, Andreas Bulling*, Andreas Breitenmoser† and Gerhard Tröster*
*Wearable Computing Laboratory, ETH Zurich, {lombriser,bulling,troester}@ife.ee.ethz.ch
†Autonomous Systems Laboratory, ETH Zurich, andreas.breitenmoser@mavt.ethz.ch

***Abstract*—One part of the vision of ubiquitous computing is the integration of sensing and actuation nodes into everyday objects, clothes worn on the body, and in large numbers into the environment. These augmented environments require novel types of interfaces that provide for naturalistic and adaptive interaction depending on user context. In this paper, we investigate the use of speech synthesis on sensor nodes that may be integrated into smart objects. We evaluate the so-called Wireless Voice Node, a small, wireless sensor node with the ability to generate voice output as a novel feedback modality for applications for ambient intelligence. As an example, we present the design and implementation of a speaking doll integrating this node. Using voice output from the doll we aim at using speech synthesis to improve the playing experience of children.**

I. Introduction

Within the vision of pervasive and ubiquitous computing, smart sensing and actuation nodes are an integral part of ambient intelligence. In the future, nodes embedded in different types of objects, on the body or in the environment are likely to implement a major part of this vision. With the increasing integration of different types of computing paradigms, unobtrusiveness of interaction and usability will become particularly challenging. One possible solution to this is to make the interfaces context-aware and use communication channels that allow for natural and seamless interaction with the user.

A common solution for interaction within smart environments is to use different characteristics of body motion such as hand or arm gestures. These gestures can be recognised using sensors integrated into body-worn items, such as the Gesture Watches [1], or into items carried along, such as the XWand [2]. Both devices can be used to interact with ambient systems by waving the hands in specific patterns. For rapid prototyping of activity-recognition systems, different toolboxes are available, which have been successfully used for various application domains (see [3] for an example).

Beigl *et al.* propose in the MediaCups project [4] to equip the objects themselves with sensors and communication capabilities. This idea can reach as far as to let people interact with virtual characters using their real-world counterpart [5].

Providing feedback from the smart object to the user is a more intricate task. Displays require significant space and power. Flashing lights or vibration allows for low-power implementation but the information that can be provided is quite limited. Speech is a natural way of interaction and has proven to provide a rich output modality, e.g. for robots [6]. However, speech usually requires a significant amount of processing for synthesis and interpretation. Software for speech synthesis on PDA class devices are available, however, these devices are rather big and cannot be integrated into small objects. Other, smaller devices only provide a preconfigured set of sounds and are too limited in the range of their expressive power.

The specific contributions of this work are (1) the evaluation and characterisation of a small low-power wireless sensor node providing speech synthesis, (2) a discussion of possible application in smart environments, smart clothes, or smart objects, and (3) the development of a speaking doll as an application for speech synthesis to enhance the playing experience of small children.

II. Wireless Voice Node

To evaluate a sensor and actuator node with speech synthesis, we have built the *Wireless Voice Node*. Its architecture is shown in Figure 1. The node contains a MSP430F1611 16bit-microcontroller running at 8MHz and a CC2420 transceiver implementing IEEE 802.15.4. As sensors we have added an ADXL330 3-axis MEMS accelerometer to detect gravity and movements performed with the device. A second sensor is a MS5540B air pressure and temperature sensor functioning as altimeter module and able to detect height changes of 50cm to 1m and a temperature range from -45° to +85°. A third APDS-9003 sensor measures ambient light and has been calibrated for indoor usage. These components can be found on standard wireless sensor nodes.

The novel component is the actuator of the Wireless Voice Node which consists of a DoubleTalk RC8660 voice synthesiser and an audio subsystem. The voice synthesiser features an integrated text-to-speech processor, a three-voice musical tone generator, and playback of up to 33 minutes of sound files stored in its onboard 7.5MB non-volatile memory.

The text-to-speech processor converts English ASCII text to speech. Different speech aspects can be adjusted, such as speed, volume, or pitch. These parameters allow altering the voice and give the smart object containing the Wireless Voice Node a personality [7]. A further useful feature of the RC8660 is its tone generator, which can produce up to three tones simultaneously over a four-octave range.

We have built a prototype module to evaluate the characteristics of the Wireless Voice Node in terms of power and its applicability for the integration into smart objects. Therefore, we have used a voice synthesiser module which includes the RC8660 as well as an audio subsystem required to play over a 8Ohm loudspeaker. The prototype has a size of 35 × 41 × 19 mm without the battery. For a future module, this size can be further reduced by using a custom implementation of

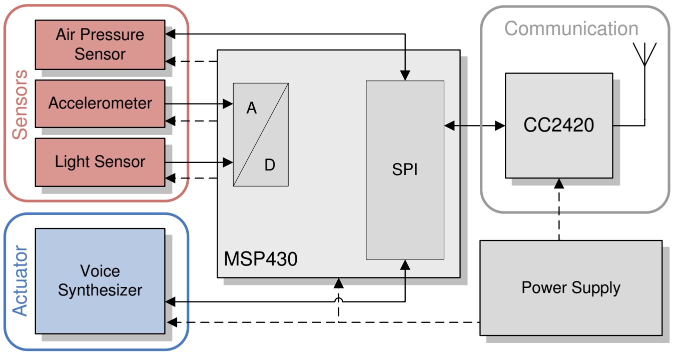

Fig. 1. The Wireless Voice Node architecture shows the 3 sensors and the voice synthesis actuator connected to the standard wireless sensor node components.

the RC8660 instead of using the voice synthesiser module. Additionally, the operating system could be ported to the DSP to get rid of the microcontroller.

Table I shows the power consumption of the Wireless Voice Node for seven different configurations. In the first four rows, the voice output is not needed, and the RC8660 is kept in suspend mode. The power values are given for sampling the acceleration sensor at 10Hz, a continuously computing MSP430 micro-controller, and in communication mode. The power consumed while producing sound is given for generating a continuous 440Hz tone, synthesising the sentence “Hello, what is your name?”, and playing back the same sentence from a file stored in the module’s flash memory.

Wireless Voice Node		
Configuration	Measured Values	
	I [mA]	P [mW]
Sampling @10Hz	1.45	5.37
MSP430 continuous	3.20	11.84
CC2420 Sending	20.35	75.30
CC2420 Receiving	20.15	74.56
440Hz output	176.92	654.62
Speech synthesis	138.78	513.48
Flash playback	204.58	756.94

TABLE I
WIRELESS VOICE NODE POWER CONSUMPTION

The measurement results clearly show that voice output is an expensive process, accounting for 90% of the power consumption when turned on. The main consumer is the amplifier circuit that produces the sound. However, also the flash memory consumes considerable power. The average power consumption is increased by 32% when using playing back the same sentence from memory instead of using speech synthesis. The table shows the power consumption when the output power of the speech synthesis is at its maximum. In the lowest output power mode, the sound is barely audible, but also consumes only 51.8% of the maximum power consumption. The intensity of the voice output perceived by a person depends on the loudspeaker and its integration into the smart object. In our demonstrator, the maximum output power produced a voice which was easily understandable in an office where other people were talking.

The device is powered by a Varta PLF443441C battery fitting the dimensions of the device. The battery provides 610mAh of typical capacity and supports an appropriate discharge current for the speech synthesiser module. This allows for approximately 42 minutes of continuous sound output or 3.8 days of continuous sampling.

Using voice feedback is a power-intensive task when compared to other functionalities of Wireless Sensor Network nodes. However, the expressive power of the feedback modality and the fact that power still can be saved when the sound output is suspended make up for the cost.

The main advantage of using a voice synthesiser is its ability to read ASCII text, such that a compact representation of speech is achieved. The text to be synthesised can be composed within the network of smart objects, where it is be exchanged in a highly compressed way. The “Hello, what is your name?” sentence needs 26 bytes encoded as a string, while it has 6074 bytes when stored in a compressed RC8660 PCM sound file. Another possibility would be to store a set of sentences or a vocabulary on the Wireless Voice Node and only indicate the file identifier required for playback. This option can be used for special sounds that may be described by a fixed string. However, if used exclusively, this would restrict the range of information that can be presented.

III. Applications

The special capabilities of the Wireless Voice Node enable new applications including networks of Wireless Voice Nodes deployed into emergency or remote areas, modules embedded into smart objects or integrated into smart clothes worn on the body. All of these applications impose similar requirements on embedded electronic devices, such as unobtrusive integration, long runtimes, or context-aware and natural interaction with the user.

A. *Smart Environments*

In case of emergency, e.g. after an earth quake or some other natural disaster, communication infrastructure often breaks down, and a potentially large number of people may be lost and unsure about what to do. In such a case a network of Wireless Voice Nodes can quickly be deployed in the emergency zone from the air. This network can not only measure different environmental parameters but is also able to give people instructions or let them trigger alarms, e.g. by shaking the device. Using voice synthesis on these nodes allows to efficiently distribute up to date emergency information over this particular area. Georouting algorithms for wireless sensor networks may also allow to lead victims to safe areas or inversely guide rescue teams to victims. The environmental sensors available on the device may help identifying potential dangers on the path, such as fire.

Displays as an alternative feedback system require direct line of sight and have difficulties in getting user attention at daylight. If eye sight is limited due to fog, smoke or dust in the air, if people are blinded or something covers access to the node, voice can be the only choice for a guidance system.

Buzzers to attract attention can only deliver limited additional information. Additionally, continuous buzzing noises could increase the stress of the people, while calm voices giving instructions may do a better job. As sound becomes useless in loud environments, a combination of the different modalities is the best choice.

B. Smart Clothes

A growing field of research are smart clothes worn on the body in daily life. Eventually, very small and light-weight nodes could be integrated into ordinary clothing such as t-shirts or head caps. For sports, Wireless Voice Nodes could be integrated into a shirt and additional sensors into the required equipment to provide a speaking sports training assistant. Such a body sensor network could e.g. monitor exercises in Tai Chi [8] or swimming [9]. The system can propose technical improvements or explain and correct training lessons targeted to improve the athlete's abilities. Using voice as feedback modality, the athlete does not have to concentrate on a screen in the vicinity and does not have to interrupt his training.

C. Smart Objects

A lot of work on smart objects has focused on how ordinary objects used in daily routine can be enriched with sensors [10]. The goal is to give these objects "intelligence" and allow for smart behaviour based on their current internal and external context. Different smart toys which are able to interact with children using audio are available on the market. However, these toys have limited context-awareness focused on locally sensed interaction, and do not interact with other toys present in the near surrounding of the child. Moreover, audio feedback provided by these toys is usually limited to short sequences of sounds or a small set of predefined words. In the future, embedded Wireless Voice Nodes would allow for context-sensitive reaction and feedback resulting in a more naturalistic way of playing.

Smart and context-dependent behaviour could not only be implemented for toys but may also change the way we use tools to build things. People lacking experience regularly face problems while trying to solve technical tasks such as assembling furniture or fixings their car. Today's manuals are provided on paper and often fail in providing the information needed to successfully perform and complete such a task. Usually there are instructions that can hardly be described only using pictures and symbols. A Wireless Voice Node provided with the package of furniture parts could provide a natural way of explaining the steps required to build up the furniture. A more complex solution would include sensors recognizing gestures on the arms, integrated in tools, or in furniture parts [11]. The nodes could connect to other nodes in the environment to improve detection of physical activity, such as hand gestures for screwing or hammering [12], and provide just-in-time audio instructions depending on the current user activity. The tools thus "know" how they should be handled and can instruct the user how to do something right.

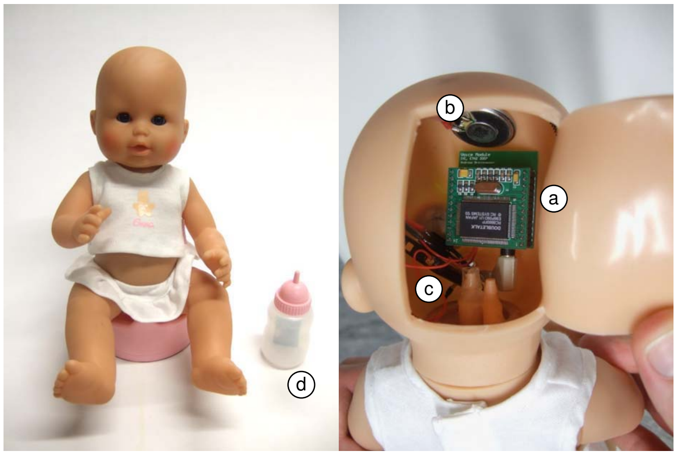

Fig. 2. Integration of the Wireless Voice Node into the speaking doll: a) the Wireless Voice Node, b) loudspeaker, c) battery, and d) a milk bottle as smart toy

IV. The Speaking Doll

To demonstrate the abilities of the Wireless Voice Node, we have integrated it into a smart toy for children: the speaking doll. In [13] it was shown that audio feedback generates an enjoyable level of interactivity with toys for children of ages between two and six. The intention of the current work was to diversify the playing experience of children by motivating them to use other, wirelessly connected smart toys they find in their environment. When the children then interact with those toys, the doll will speak a sentence related to the actions.

A. Intended Interaction

When a child is playing with the doll, the doll will ask for another smart toy in the environment from time to time. Each smart toy identifies itself to the doll with a string representing its human-readable name. The doll can read this string to the child and ask to play with it. For this purpose, the smart toys send a set of verbs specifying actions that can be performed with themselves. The doll can thus produce a sentence like "I want *drink* `milk`" if it perceives a smart milk bottle in its vicinity.

Using algorithms for context recognition, the smart toys can e.g. recognise motions that are performed with them [14]. In case they detect a specific motion, they can send a string to the doll, which will give the player a feedback to their action. This could be drinking sounds when the smart milk bottle is tilted.

The action sentences are only played when the smart objects are in the immediate proximity of the doll. If they are not close, the doll assumes that they are being used in conjunction with other smart objects and the sentence sent is considered irrelevant.

It is important to note that all the information concerning the identity or the actions on the smart objects is provided by the smart objects themselves. The doll does not keep any information about what smart toys it can interact with. This has the advantage that the doll can interact with smart

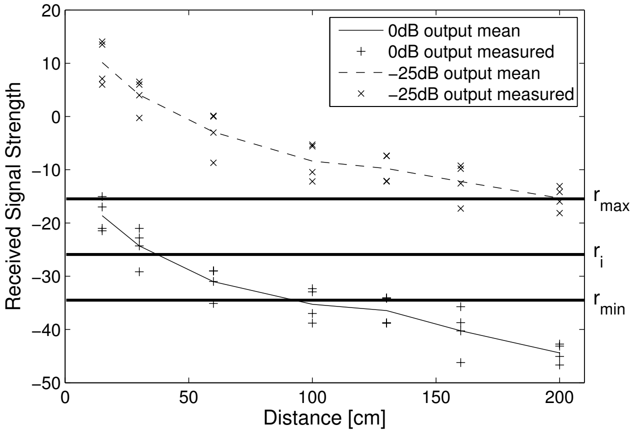

Fig. 3. RSSI measurements around the doll. Distance thresholds for determining the smart object proximity are given for -25dB output power.

toys that were not known at design time. Also, the play is completely independent of any infrastructure and can be performed anywhere.

B. Implementation

In our implementation, the smart objects are programmed to periodically send identification messages with their human readable names and associated verbs to the doll. Further, the smart objects process their local sensor readings and issue context messages in text form in case they recognise some activity or situation they are trained for.

Upon the reception of a *context message*, the Wireless Voice Node uses a Received Signal Strength Indication (RSSI) of the message r to determine whether a smart object is close enough to the doll. Fig. 3 shows our RSSI measurements in the environment of the doll for maximum and minimum sender output power. The lines present average values of RSSI measurements of messages from smart objects in front, back, and to both sides of the doll. An inverse square dependency of the RSSI on the distance can clearly be seen. This dependency can be used to set a threshold for determining whether a smart object is in the immediate proximity ($r_{max} \geq r \geq r_i$), close proximity ($r_i > r \geq r_{min}$), or out of reach ($r_{min} > r$). In immediate proximity, the recognised context delivered by a smart object is considered relevant and will be played by the doll.

The challenges in collecting proximity data for interacting wireless sensor nodes are described by experiments by [15]. In our approach we profit from the steep increase in RSSI in close proximity, which makes the evaluation more reliable.

Upon reception of a smart object's *identification message*, the doll uses a probability value as criterion for acceptance. This probability is influenced by two factors: (1) whether the player has interacted with the smart object in recent time, and (2) how close the object is to the doll. The proximity is again determined using RSSI and is used to scale the acceptance probability with the distance between smart toy and doll. If the RSSI value r is smaller than the out-of-reach threshold r_{min}, the identification message is dropped.

To remember interactions with the player, the doll remembers the wireless network node address of the smart object and keeps a value t_{used} counting how many times context messages have been received from this particular smart object. It is assumed that context messages are only issued by a smart object in case that the player has somehow interacted with it. The t_{used} count is periodically decremented, such that only a certain time frame of possible interactions is considered.

The probability for acceptance of the identification messages follows to be:

$$p_{acc}(t_{used}, r) \quad = \quad \frac{r_{max} - r}{r_{max} - r_{min}} \frac{1}{1 + t_{used}} p_{max} \qquad (1)$$

In the time that t_{used} and r do not change, the mean time t_{acc} until an identification message is accepted can be modeled using a geometric distribution. It can be influenced by setting p_{max} and follows the following relation:

$$\mathrm{E}[t_{acc}] \quad = \quad \frac{r_{max} - r_{min}}{r_{max} - r} \frac{1 + t_{used}}{p_{max}} \qquad (2)$$

Meaningful values for p_{max} need to be determined by actually letting children play with the doll. This is subject of future work. For development, the values of Figure 3 can be used as a reference.

C. Speech Generation

The speaking doll composes sentences from identification messages using templates containing placeholders for smart toy names and verbs. Additionally, the doll stores a set of sounds such as laughing, crying, drinking, and burping sounds. Using escape values, these sounds can be addressed within the sentences sent by the smart toys.

The advantage of using a doll is that it does not require to generate grammatically correct sentences. Thus the speech generation can be kept simple and targeted to the main objective: to augment and diversify playing with the doll. Speech recognition and generation are issues which research in artificial intelligence and robotics have been concerned with for a considerable time. [16] gives an overview over the state of the art, and identifies a limitation in speech processing in missing flexibility and adaptation to the user's context. Our approach to the speaking doll enables talking devices to expand their vocabulary dynamically whenever new nodes are in the proximity, and delivers speech generated timely and relevant to the user's context.

V. Discussion

Using speech as a feedback modality has the the advantage of hiding the technology from the human user. Neither the hands nor the eyes need to be fixed on some interface device, giving the user more freedom in his activities. The applications we have described especially point out that using speech, the technology is not directly visible to the user and thus lets him concentrate on other tasks he is up to at the moment. Of course, speech may also be used in conjunction with other feedback modalities, but also has a rich information content on its own, as it can also implicitly communicate emotions through modification of the voice's personality [7].

The characterization of the Wireless Voice Node shows that the integration of speech synthesis is indeed feasible for smart objects. The power consumption is high, but the module can be suspended when it is not needed. Representation of speech as ASCII strings further allows exchanging voice messages in a very compact way. Sounds that are hard to express by text can be stored on the Wireless Voice Node and called by special codes in the text when necessary.

The speaking doll is an example application for speech generation using a Wireless Voice Node. The speaking doll has the ability to integrate previously unknown smart toys into the game. The speech output is generated at runtime and addresses specific actions performed by the player, even if no information about the smart toys was available at design time of the speaking doll.

The intelligence required to run the game is distributed to the smart objects taking part. Each smart object recognizes player actions associated with itself and is able to express them via the Wireless Voice Node in the doll. No additional management functionality is needed for distributing the logic. For distributed processing and fusion of recognition results, the speech functionality could be integrated into frameworks for distributed processing such as Titan [17].

In future work, we want to conduct user studies with children in which we plan to evaluate if additional interest in playing with the toys using our feedback application can be achieved as described for a similar application [13]. An interesting question will be if and how different voice personalities will have influence in capturing the children's interest.

A further subject of research will be the distributed generation of spoken sentences. Multiple smart objects will collaboratively generate sentences to be spoken by the Wireless Voice Node. In a possible application, the speaking doll could become jealous if the player plays for too long with two other toys.

VI. Conclusion

In this paper we have proposed a novel modality for interaction with smart objects. We have described a speech synthesis module that enables sensor and actuator nodes to use a natural way of interaction with persons using the system. We have characterized a prototype and presented possible applications, which can benefit from using speech as an output modality. On the example of a speaking doll we described how such an application can be designed and implemented. The example demonstrates how speech can be used as a feedback modality to enhance the gameplay experience, and is a fully distributed application for smart objects.

Acknowledgments

This paper describes work undertaken in the context of the SENSEI project, 'Integrating the Physical with the Digital World of the Network of the Future' (www.sensei-project.eu). SENSEI is a Large Scale Collaborative Project supported by the European 7th Framework Programme, contract number: 215923. The authors also would like to thank the students of the "Smart Toys" project of the summer semester 2008 for their ideas for pervasive games.

References

[1] J. Kim, J. He, K. Lyons, and T. Starner, "The gesture watch: A wireless contact-free gesture based wrist interface," in *Proc. 11th Int. Symp. Wearable Computers (ISWC)*, pp. 15–22, 2007.

[2] A. Wilson and S. Shafer, "Xwand: Ui for intelligent spaces," in *Proc. SIGCHI Conf. on Human Factors in Computing Systems*, pp. 545–552, 2003.

[3] D. Bannach, O. Amft, and P. Lukowicz, "Rapid prototyping of activity recognition applications," *IEEE Pervasive Computing*, vol. 7, no. 2, pp. 22–31, 2008.

[4] M. Beigl, H. Gellersen, and A. Schmidt, "MediaCups: Experience with design and use of computer-augmented everyday artefacts," *Computer Networks*, vol. 35, pp. 401–409, 2001.

[5] M. P. Johnson, A. Wilson, B. Blumberg, C. Kline, and A. Bobick, "Sympathetic interfaces: using a plush toy to direct synthetic characters," in *Proc. SIGCHI Conf. Human Factors in Computing Systems*, pp. 152–158, 1999.

[6] C. Breazeal and L. Aryananda, "Recognition of affective communicative intent in robot-directed speech," *Autonomous Robots*, vol. 12, pp. 83–104, 2002.

[7] M. Schmitz, A. Krüger, and S. Schmidt, "Modelling personality in voices of talking products through prosodic parameters," in *Proc. 12th Int. Conf. Intelligent User Interfaces (IUI)*, pp. 313–316, 2007.

[8] D. Majoe, I. Kulka, and J. Gutknecht, "Qi energy flow visualisation using wearable computing," in *Proc. 2nd Int. Conf. Pervasive Computing and Applications (ICPCA)*, pp. 285–290, 2007.

[9] M. Bächlin, K. Förster, J. Schumm, D. Breu, J. Germann, and G. Tröster, "An automatic parameter extraction method for the 7x50m stroke efficiency test," in *Proc. 3rd Int. Conf. Pervasive Computing and Applications (ICPCA)*, pp. 442–447, 2008.

[10] K. Römer and F. Mattern, "The design space of wireless sensor networks," *IEEE Wireless Communications*, vol. 11, pp. 54–61, 2004.

[11] S. Antifakos, F. Michahelles, and B. Schiele, "Proactive instructions for furniture assembly," in *Proceedings of the 4th International Conference on Ubiquitous Computing (UbiComp)*, pp. 351–360, 2002.

[12] P. Lukowicz, J. A. Ward, H. Junker, M. Stäger, G. Tröster, A. Atrash, and T. Starner, "Recognizing workshop activity using body worn microphones and accelerometers," in *Proc. 2nd Int. Conference on Pervasive Computing*, pp. 18–32, 2004.

[13] W. Fontijn and P. Mendels, "Storytoy the interactive storytelling toy," in *Proc. 2nd Int. Workshop on Gaming Applications in Pervasive Computing Environments (PerGames)*, 2005.

[14] C. Lombriser, N. B. Bharatula, D. Rogen, and G. Tröster, "On-body activity recognition in a dynamic sensor network," in *Proc. 2nd Int. Conf. Body Area Networks (BodyNets)*, 2007.

[15] W. Brunette, C. Hartung, B. Nordstrom, and G. Borriello, "Proximity interactions between wireless sensors and their application," in *Proc. 2nd ACM Int. Conf. Wireless Sensor Networks and Applications (WSNA)*, pp. 30–37, 2003.

[16] R. Prasad, H. Saruwatari, and K. Shikano, "Robots that can hear, understand and talk," *Advanced Robotics*, vol. 18, no. 5, pp. 533–564, 2004.

[17] C. Lombriser, D. Roggen, M. Stäger, and G. Tröster, "Titan: A tiny task network for dynamically reconfigurable heterogeneous sensor networks," in *Proc. 15. Fachtagung Kommunikation in Verteilten Systemen (KiVS)*, pp. 127–138, 2007.